A MISSING INFORMATION PRINCIPLE: THEORY AND APPLICATIONS

TERENCE ORCHARD
and
MAX A. WOODBURY
DUKE UNIVERSITY MEDICAL CENTER

1. Introduction

The problem that a relatively simple analysis is changed into a complex one just because some of the information is missing, is one which faces most practicing statisticians at some point in their career. Obviously the best way to treat missing information problems is not to have them. Unfortunately circumstances arise in which information is missing and nothing can be done to replace it for one reason or another. In analogy to other accidents—we don't plan on accidents, nevertheless they do occur and safety measures must be aimed at palliating consequences as well as at prevention. Consequently, a great volume of literature has been produced, dealing with a number of specific situations. An indication of the content of many of these papers is given in the Appendix. In this paper we propose to try to present a general philosophy for dealing with the problem of missing information, and to give a method which will lead quite easily to maximum likelihood estimates of the parameters obtained from the incomplete data using as nearly as possible the same techniques as if the data were all present.

Our first simple use of the missing information principle resulted from a conversation in 1946 between Max A. Woodbury and C. W. Cotterman resulting from the latter's interest (Cotterman [20]) in estimating gene frequencies from phenotypic frequency data. The observation was made that if one has the genotypic *frequencies* NAA, NAB, NBA, NBB, NAO, NBO, NOB, and NOO of red blood cell genotypes, indicated by the second and third letters of each symbol, then the gene frequencies are easily computed. If N is the total of the above frequencies then the estimates would be

$$\begin{aligned}\hat{p}_A &= \frac{1}{2N}(2NAA + NAB + NBA + NAO + NOA),\\ \hat{p}_B &= \frac{1}{2N}(2NBB + NBA + NAB + NBO + NOB),\\ \hat{p}_o &= \frac{1}{2N}(2NOO + NOA + NAO + NOB + NBO).\end{aligned} \tag{1.1}$$

Research conducted under PHS-NIH Grant GM 16725-02 from the National Institute of General Medical Sciences.

However, only the phenotypic frequencies

$$\begin{aligned} MA &= NAA + NAO + NOA, \\ MB &= NBB + NBO + NOB, \\ MO &= NOO, \\ MAB &= NAB + NBA, \end{aligned} \tag{1.2}$$

are available.

If, however, one makes use of Bayes' theorem and the gene frequencies one can obtain estimates of the genotypic frequencies from the phenotypic frequencies

$$\begin{aligned} N\hat{A}A &= MA\,\frac{p_A^2}{(p_A^2 + 2p_Ap_O)} = MA\,\frac{p_A}{(p_A + 2p_O)}, \\ N\hat{A}O + N\hat{O}A &= 2MA\,\frac{p_A}{(p_A + 2p_O)}, \\ N\hat{B}B &= MB\,\frac{p_B}{(p_B + 2p_O)}, \\ N\hat{B}O + N\hat{O}B &= 2MB\,\frac{p_O}{(p_B + 2p_O)}, \\ N\hat{A}B + N\hat{B}A &= MAB, \\ N\hat{O}O &= MO. \end{aligned} \tag{1.3}$$

If one solves (1.1) and (1.3) simultaneously by equating the genotypic frequencies in (1.1) to their estimates in (1.3), one can obtain estimates $\tilde{p}_A$, $\tilde{p}_B$, and $\tilde{p}_O$, which of course are not as good as those obtainable from the true genotype frequencies but which are as efficient as the maximum likelihood estimates based only on the phenotypic frequencies.

The problem with estimating p_A, p_B, and p_O by this method is the difficulty of finding rapid and accurate solutions of these equations and estimates of their error variances. These difficulties are shared with the method of Maximum Likelihood (ML). This is not too surprising since in fact the two methods are equivalent. One way in which the two problems of slow convergence and loss of information may be handled is by the method of scoring which can be modified to work in the presence of missing information. The solution of the problem of estimating gene frequencies in more general circumstances is provided by Ceppellini, Siniscalco, and Smith [15], who demonstrated that the procedure implied by the principle indicated above is in fact ML in all cases. These authors also considered the increased variance of the estimates due to the loss of genotypic information under the heading "hidden variance."

This missing information principle has been applied to missing observations in a linear model, and in a multivariate normal, and to mixture problems. A few examples will be presented later in order to demonstrate the relative facility with which the principle may be applied.

2. Theory

The method proposed is to regard the values of the missing data as random variables within the framework of a model of the data. Thus, estimates which are well defined when all the data are present become random variables (being functions of the missing data). This variation of the estimates is in addition to the usual sampling variation so that the error variances of the estimates are increased. The consequences of the data's being missing and some insight to the approach used here are obtained if one considers replacing the missing data by sample values from the appropriate distribution function. The question is, from which distribution function should we sample?

In the independent, identically distributed case where the vector x_i has the distribution function $f(x_i|\theta)$ and $x_i = (Y_i, z_i)$ where the vector z_i contains the missing components, then $f(x_i|\theta) = f_1(y_i|\theta)\cdot f_2(z_i|\theta, y_i)$ is the factorization of the distribution function into the marginal distribution for y_i and the conditional distribution of z_i given y_i. The proper distribution to sample for the missing data then is the conditional distribution $f(z_i|\theta, y_i)$, but θ is unknown so that some estimated value $\hat{\theta}$, must be used. One could draw many samples from the distribution f and from these completed data samples obtain the distribution of the parameter estimates due to the missing data. Call this distribution MID$(\hat{\theta}, Z)$. If this distribution is asymptotically normal, then the mean will be the obvious statistic to use to provide an estimate in the presence of missing data. If this mean value should be $\hat{\theta}$, then the estimate has not been affected by the assumed missing data distribution. That is the missing data tells you nothing. This interpretation of the principle is due to Jacquez. The remaining part of the missing information principle is to equate the mean of the MID$(\hat{\theta}, Z)$ to $\hat{\theta}$, or take some action equivalent to this. The effect of the variance of MID$(\hat{\theta}, Z)$ ("the hidden variances") on the error variance is best understood in another context.

Before continuing with the problem of missing information, it will be of value to review the method of maximum likelihood estimation. The likelihood function of a multivariate data matrix X is denoted by $L(X|\theta)$, and is defined to be $L(X|\theta) = \Pi_{n=1}^{N} f(X_n|\theta)$, where $f(X_n|\theta)$ is the density function of X_n and θ is an $(s \times 1)$ vector of parameters. The Score (Sco) for the parameter θ is then defined to be

$$\text{Sco}(\theta_j|X) = \frac{\partial}{\partial\theta_j} \log L(X|\theta) = \frac{1}{L}\frac{\partial}{\partial\theta_j} L(X|\theta). \tag{2.1}$$

The maximum likelihood estimates for the parameters are obtained by solving the set of equations. It may be readily deduced that the mean value of the score is zero at the true parameter point; that is,

$$\text{Sco}\,(\theta_j|X) = 0 \qquad \text{for} \quad j = 1, \cdots, s, \tag{2.2}$$

$$E[\text{Sco}(\hat{\theta}_j|X)] = 0. \tag{2.3}$$

The information matrix for θ is defined to be the matrix with (j, k)th element

$$
\begin{aligned}
(2.4) \qquad J(\theta_j, \theta_k|X) &= -E\left[\frac{\partial}{\partial\theta_k}\operatorname{Sco}(\theta_j|X)\right] \\
&= -E\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log L(X|\theta)\right] \\
&= E[\operatorname{Sco}(\theta_j|X)\operatorname{Sco}(\theta_k|X)] \\
&= \operatorname{Cov}[\operatorname{Sco}(\theta_j|X), \operatorname{Sco}(\theta_k|X)]
\end{aligned}
$$

Under certain general regularity conditions (see Rao [36], p. 295), for example, the existence of second and third derivatives of the log likelihood function, it may be shown that the joint distribution of the maximum likelihood estimators is asymptotically multivariate normal with a covariance matrix given by the inverse of the information matrix. For a detailed discussion of this subject, the paper by Wald [41] can be consulted. A condition for the existence of maximum likelihood estimates is that the information matrix is positive definite. However, in some instances, such as in the case of the multinomial distribution, the information matrix has rank less than s and is therefore singular. If the information matrix is of rank $s - t$, then we are required to impose t restrictions, $h_1(\theta) = 0, \cdots, h_t(\theta) = 0$, on the parameters in order to achieve identifiability.

The problem of maximum likelihood estimation of parameters subject to constraints has been studied by Aitchison and Silvey [5], [6], and Silvey [38]. These authors also obtained a test statistic for the hypothesis $h(\theta) = 0$. However, the situation of interest here is when $h(\theta)$ is a $t \times 1$ vector of constraints necessary to produce an identifiable parameter set. In this case the constrained likelihood function may be written as

$$
(2.5) \qquad L^* = \log L(X|\theta) - \lambda^T h(\theta),
$$

where λ is a $t \times 1$ vector of Lagrangian multipliers. If we define the constrained score

$$
(2.6) \qquad \operatorname{Sco}^*(\theta|X) = \frac{\partial L^*}{\partial\theta}
$$

and

$$
(2.7) \qquad \underset{(s\times t)}{H} = \left(\frac{\partial h_i(\theta)}{\partial\theta_j}\right),
$$

then we may deduce that the expected value of the constrained score is zero if the true parameter satisfies the constraints.

Considering the various definitions of the information matrix listed in (2.4), we may determine that a nonsingular matrix, denoted by J^*, will be obtained by taking the negative of the expected value of the derivative of the constrained scores. However, if we take the covariance of the constrained scores we will call

the singular information matrix J. This is related to the required nonsingular matrix by the equation

$$J^* = J + NHH^T, \tag{2.8}$$

where N is the number of observations. This form will lead to the asymptotic covariance matrix of the parameter estimates given by

$$V = (J^*)^{-1} = (J^{-1} - J^{-1}H(H^TJ^{-1}H)^{-1}H^TJ^{-1}). \tag{2.9}$$

It may also be remarked that a nonsingular information matrix will not result simply from reparametrizing so that the new parameters satisfy the constraints, unless some are eliminated.

Let us now return to the situation where we have missing information. Suppose that we cannot observe the random variable X, but can instead observe some image Y, of it. The likelihood function for Y may be obtained by integrating $L(X|\theta)$ over the appropriate range, and thus we may write

$$L(X|\theta) = L_1(X|Y, \theta)\, L_2(Y|\theta) \tag{2.10}$$

giving

$$\text{Sco}\,(\theta_j|X) = \frac{1}{L_1}\frac{\partial L_1}{\partial \theta_j} + \text{Sco}\,(\theta_j|Y). \tag{2.11}$$

The item $(1/L_1)\,(\partial L_1/\partial\theta_j)$ is called the conditional score of θ_j from X, given Y, and this is denoted by Sco $(\theta_j, X|Y)$. It may be noted that

$$E[\text{Sco}\,(\theta_j, X|Y)|Y] = 0, \tag{2.12}$$

the truth of this following from the same reasoning as was used to establish (2.3).

Finally we have

$$\text{Sco}\,(\theta_j|Y) = E[\text{Sco}\,(\theta_j|X)|Y] \tag{2.13}$$

and

$$E[\text{Sco}\,(\theta_j|X)\,\text{Sco}\,(\theta_k|X)] = \text{Cov}\,\{E[\text{Sco}\,(\theta_j|X)|Y], E[\text{Sco}\,(\theta_k|X)|Y]\} + E\{\text{Cov}\,[\text{Sco}\,(\theta_j|X), \text{Sco}\,(\theta_k|X)|Y]\} \tag{2.14}$$

which leads to

$$E[\text{Sco}\,(\theta_j|X)\,\text{Sco}\,(\theta_k|X)] = E[\text{Sco}\,(\theta_j|Y)\,\text{Sco}\,(\theta_k|Y)] + E\{\text{Cov}\,[\text{Sco}\,(\theta_j|X), \text{Sco}\,(\theta_k|X)|Y]\}. \tag{2.15}$$

For the information matrix this gives

$$J(\theta, \theta|X) = J(\theta, \theta|Y) + J(\theta, \theta; Y|X). \tag{2.16}$$

The last quantity on the right $J(\theta, \theta; Y|X)$ is what is termed the lost information. For brevity, equation (2.16) may be written $J_X = J_Y + J_{X/Y}$, where $J_{X/Y} = J_X - J_Y$ = the lost information.

We may now easily obtain a relationship between the lost information and the increase in variance of the parameter estimates (the "hidden variance" of Ceppellini, Siniscalco, and Smith [15]). This relationship derives from

$$J_X(J_Y^{-1} - J_X^{-1}) = (J_X - J_Y)J_Y^{-1} \tag{2.17}$$

which we may write as

$$J_Y^{-1} - J_X^{-1} = J_X^{-1}(J_X - J_Y)J_Y^{-1}, \tag{2.18}$$

$$J_X - J_Y = J_X(J_Y^{-1} - J_X^{-1})J_Y. \tag{2.19}$$

If, for simplicity, we write $A \geqq B$ when the matrix difference, $C = A - B$, is positive semidefinite, then we have

$$J_Y^{-1} \geqq J_X^{-1}, \tag{2.20}$$

$$J_X \geqq J_Y. \tag{2.21}$$

We may now obtain bounds on the hidden variance in terms of the lost information, and vice versa. These are

$$J_X^{-1}J_{X|Y}J_X^{-1} \leqq (J_Y^{-1} - J_X^{-1}) \leqq J_Y^{-1}J_{X|Y}J_Y^{-1}. \tag{2.22}$$

$$J_Y(J_Y^{-1} - J_X^{-1})J_Y \leqq J_{X|Y} \leqq J_X(J_Y^{-1} - J_X^{-1})J_X. \tag{2.23}$$

These may be of value in practical situations where some of the quantities are easier to obtain than others. The widths of these limits depend on the amount of missing information and are "tight" when this is small.

The usefulness of the above theory, and in particular result (2.13), in estimating parameters in the presence of missing information is that it is often quite easy to obtain the right side even in those cases when it is extremely difficult to obtain the left side.

3. Examples

3.1. *Example 1.* Consider first the case of missing observations in a linear model. Suppose that Y is a set of independent, normally distributed, random variables having a common variance of σ^2, and a mean of $X\theta$, where X is an $n \times k$ design matrix and θ is a $k \times 1$ vector of unknown parameters. Then we have

$$\text{Sco}\,(\theta|Y) = \frac{X^T(Y - X\theta)}{\sigma^2}, \tag{3.1.1}$$

$$\text{Sco}\,(\sigma^2|Y) = \frac{n}{2\sigma^2} - \frac{(Y - X\theta)^T(Y - X\theta)}{2\sigma^4}. \tag{3.1.2}$$

Equating these to zero gives

$$\hat{\theta} = (X^TX)^{-1}X^TY, \tag{3.1.3}$$

$$\hat{\sigma}^2 = \frac{1}{n}(Y - X\theta)^T(Y - X\theta) \tag{3.1.4}$$

This estimator for σ^2 is, of course, well known to be biased, the unbiased estimator being

$$\hat{\sigma}^2 = \frac{(Y - X\theta)^T(Y - X\theta)}{n - k}, \tag{3.1.5}$$

where k is the rank of the design matrix X.

Suppose that there are n potential observations but that there are m missing, leading to an image of Y being the vector Y_0 of observed values. Due to the assumed independence of the observations the conditional expectation of the scores may be easily computed. Also the observed values are unaltered whilst functions of the missing values are replaced by their expected values. Hence

$$\text{Sco}\,(\theta|Y_0) = \frac{1}{\sigma^2} X^T(Y_0 - X_m\theta - X\theta) \tag{3.1.6}$$

and

$$\text{Sco}\,(\sigma^2|Y_0) = \frac{n}{2\sigma^2} - \frac{(Y_0 - X_0\theta)^T(Y_0 - X_0\theta)}{2\sigma^2} - \frac{m}{2\sigma^2}; \tag{3.1.7}$$

here X_m and X_0 are $(n \times k)$ matrices such that $X = X_m + X_0$, and Y_m and Y_0 are $(n \times 1)$ vectors such that $Y = Y_m + Y_0$. The following estimators are thus obtained:

$$\hat{\theta} = (X^TX)^{-1}X^T\hat{Y}, \tag{3.1.8}$$

$$\hat{Y} = Y_0 + X_m\hat{\theta}, \tag{3.1.9}$$

$$\hat{\sigma}^2 = (Y_0 - X_0\hat{\theta})^T \frac{(Y_0 - X_0\hat{\theta})}{n - m - k} \tag{3.1.10}$$

$$= (\hat{Y} - X\hat{\theta})^T \frac{(\hat{Y} - X\hat{\theta})}{n - m - k}.$$

It may be noted that $\hat{\theta}$ can be eliminated between (3.1.8) and (3.1.9) to obtain

$$\hat{Y}_m = [I - X_m(X^TX)^{-1}X_m^T]^{-1}X_m(X^TX)^{-1}X_0^TY_0. \tag{3.1.11}$$

This equation provides a form of estimating missing data which is easy to compute for a single value. It is proposed to use this simple form to obtain an initial set of estimates for the missing values, and then to cycle iteratively through equations (3.1.8) and (3.1.9) until the parameter estimates stabilize. This is the modification of Yates' [50] approach to the problem, as proposed by Tocher [39].

The lost information for estimating $\hat{\theta}$ may be shown to be $(X_m^TX_m)\sigma^{-2}$, thus the covariance matrix for $\hat{\theta}$ may be written

$$\begin{aligned}\text{Cov}\,(\hat{\theta}, \hat{\theta}) &= (X^TX - X_m^TX_m)^{-1}\sigma^2 \\ &= \{(X^TX)^{-1} + (X^TX)^{-1}X_m^T[I - X_m(X^TX)^{-1}X_m^T]^{-1}X_m \\ &\qquad (X^TX)^{-1}\}\sigma^2.\end{aligned} \tag{3.1.12}$$

Therefore the quantity to be added to Cov $(\hat{\theta}, \hat{\theta})$ so as to correct for "bias" is

$$B = \{(X^TX)^{-1}X_m^T[I - X_m(X^TX)^{-1}X_m^T]^{-1}X_m(X^TX)^{-1}\}\sigma^2. \tag{3.1.13}$$

Since we recover none of the lost information, in general the main reason for using the procedure proposed here is that the design matrix X for the complete data, where we have a balanced design, may be chosen such that X^TX is diagonal, and hence much easier to invert than the general matrix $X_0^TX_0$ which would result from using the available data only. However, it should be noted that in order to correct for the bias in Var (θ), it is necessary to invert the matrix $[I - X_m^T(X^TX)^{-1}X_m]$. This is of order m and generally has a regular pattern, but it may not be diagonal. It is therefore felt that if the number of missing values m equals the rank of the design matrix then there is less to be gained from using the procedure outlined here.

3.2. *Example 2.* A second example, which has received considerable attention, is that of missing components in a multivariate normal distribution. This has been considered by Woodbury and Hasselblad [47] but it is being included here due to its great interest. The log of the likelihood function, for a sample of size N, may be written

$$\log L(X|\mu, \Sigma) = C + \tfrac{1}{2}N|\Sigma^{-1}| - \tfrac{1}{2}N \operatorname{tr} \Sigma^{-1}S, \tag{3.2.1}$$

where $S = \Sigma_{n=1}^{N}(X_n - \mu)(X_n - \mu)^T/N$.

The parameter scores are easily obtained and are

$$\operatorname{Sco}(\mu|X) = N\Sigma^{-1}(\bar{X} - \mu) \tag{3.2.2}$$

and

$$\operatorname{Sco}(\Sigma|X) = -\tfrac{1}{2}N(\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1}). \tag{3.2.3}$$

By equating these to zero we can obtain the parameter estimates

$$\hat{\Sigma} = S \tag{3.2.4}$$

and

$$\hat{\mu} = \bar{X} = \sum_{n=1}^{N} \frac{X_n}{N}. \tag{3.2.5}$$

Equation (3.2.3) is obtained by observing that the derivative of the logarithm of a determinant of a matrix with respect to an element of that matrix is simply the corresponding element of the inverse, and that $\partial\Sigma^{-1} = -\Sigma^{-1}(\partial\Sigma)\Sigma^{-1}$. We also used such properties as tr (AB) = tr (BA), and tr $(a) = a$ if a is a scalar. Additionally we note that $\sigma^{i,j} = \sigma^{j,i}$, although this fact is not used in obtaining Sco $(\sigma_{i,j})$. The solution of the normal equations will not be affected since we obviously have Sco $(\sigma_{i,j})$ = Sco $(\sigma_{j,i})$. The reason for forming the score of the covariance matrix instead of the score of the inverse, as is more normal, is that this will greatly simplify the computation of the information for Σ, as will be developed later.

The image Y of X consists of the observed components of the data matrix X. If we assume that there are p components then an individual observation is a $p \times 1$ vector which may be written in the form

$$\hat{Y}_k = Y_{k,0} + \hat{Y}_{k,m}, \tag{3.2.6}$$

where $Y_{k,0}$ is the observed portion, with zero in each position corresponding to a missing component, and $\hat{Y}_{k,m}$ is the estimated missing portion, with again zero in the positions corresponding to an observed component. It should of course be remarked that if there are no missing components then $Y_{k,0}$ constitutes the entire vector. To obtain the conditional expectation of the scores for the mean we have to solve the regression of the missing data $Y_{k,m}$ on the observed data $Y_{k,0}$. Similarly the conditional expectation of the scores for the covariance matrix requires the conditional covariance matrix of the missing data given the observed data. The following estimators are obtained:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} Y_n, \tag{3.2.7}$$

$$\hat{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} [(\hat{Y}_n - \hat{\mu})(\hat{Y}_n - \hat{\mu})^T + V_n], \tag{3.2.8}$$

$$\begin{aligned} \hat{Y}_{k,m} &= \hat{\mu}_m + \Sigma_{m,0}\Sigma_{0,0}^{-1}(Y_{k,0} - \hat{\mu}_0) \\ &= \hat{\mu}_m + (\Sigma^{m,m})^{-1}\Sigma^{m,0}(Y_{k,0} - \hat{\mu}_0), \end{aligned} \tag{3.2.9}$$

where V_n is a $p \times p$ matrix, for the nth observation, with $\Sigma_{m,m} - \Sigma_{m,0}\Sigma_{0,0}^{-1}\Sigma_{0,m}$, $(= (\Sigma^{m,m})^{-1})$, in the positions corresponding to the missing components and zero elsewhere. It should be noted that partitions such as $(Y_{k,m}, Y_{k,0})$ vary from observation to observation depending on which components, if any, are missing.

The lost information for the mean due to a single observation may be shown to be

$$LI(\mu) = \begin{bmatrix} \Sigma^{m,m} & \Sigma^{m,0} \\ \Sigma^{0,m} & \Sigma^{0,0} - \Sigma_{0,0}^{-1} \end{bmatrix} \tag{3.2.10}$$

This may be deduced intuitively since we would lose Σ^{-1} by discarding a complete observation, whereas the procedure described here recovers the amount $\Sigma_{0,0}^{-1}$ contained in the observed portion. Once again the components of $\Sigma_{0,0}$ vary from observation to observation and will be the entire matrix if no component is missing.

Since there are only $q = \frac{1}{2}p(p + 1)$ distinct elements in the covariance matrix, $J(\Sigma, \Sigma)$ will be a $q \times q$ matrix. However, it is proposed to regard it as being composed of p^2 distinct elements and then to gather the terms in the $p^2 \times p^2$ information matrix $J^*(\Sigma, \Sigma)$ to give the required $J(\Sigma, \Sigma)$. Thus we have

$$(3.2.11)\qquad J^*(\Sigma, \Sigma) = \frac{N^2}{4} E[\Sigma^{-1} \otimes \Sigma^{-1} - \Sigma^{-1} \otimes \Sigma^{-1} S \Sigma^{-1} - \Sigma^{-1} S \Sigma^{-1} \otimes \Sigma^{-1} + \Sigma^{-1} S \Sigma^{-1} \otimes \Sigma^{-1} S \Sigma^{-1}]$$
$$= \frac{N^2}{4} E[(\Sigma^{-1} \otimes \Sigma^{-1})(S \otimes S)(\Sigma^{-1} \otimes \Sigma^{-1}) - \Sigma^{-1} \otimes \Sigma^{-1}]$$

A typical element in $E(S \otimes S)$ may be shown to be

$$(3.2.12)\qquad E(s_{i,j} s_{u,v}) = \frac{1}{N} [\sigma_{i,j}\sigma_{u,v} + \sigma_{i,u}\sigma_{j,v} + \sigma_{i,v}\sigma_{j,u}]$$

Hence (3.2.11) reduces to

$$(3.2.13)\qquad J^*(\Sigma, \Sigma) = \frac{N}{4} (\Sigma^{-1} \otimes \Sigma^{-1})(\sigma_{i,u}\sigma_{v,j} + \sigma_{i,v}\sigma_{u,j})(\Sigma^{-1} \otimes \Sigma^{-1})$$
$$= \frac{N}{4} (\sigma^{i,u}\sigma^{v,j} + \sigma^{i,v}\sigma^{u,j}).$$

To obtain these equations it is necessary to use one of the properties of the Kronecker product, namely, $(A \otimes B)(C \otimes D) = AC \otimes BD$. To obtain $J(\Sigma, \Sigma)$ we simply collapse this matrix noting that $J(\sigma_{i,i}, \sigma_{u,u})$ consists of one element like that shown on the right side of (3.2.13), $J(\sigma_{i,i}, \sigma_{u,v})$ consists of the sum of two such elements, and $J(\sigma_{i,j}, \sigma_{u,v})$ consists of the sum of four such elements. The point of obtaining $J^*(\Sigma, \Sigma)$ is that it provides a means of computing the information matrix in the presence of missing data since J^* is the sum of the corresponding matrices for each observation vector, which may differ from observation to observation in the case of missing data. If we write each observation in the form $\hat{Y}_k = Y_{k,0} + \hat{Y}_{k,m}$ then $\Sigma_{0,0}$ is the submatrix of Σ corresponding to $Y_{k,0}$. The retained information is then obtained by accumulating those elements of (3.2.13) that correspond to $\Sigma_{0,0}^{-1}$. Once this has been done then J can be obtained by combining terms and from this the lost information may be computed, as can the covariance matrix of the estimated covariance matrix. covariance matrix of the estimated covariance matrix.

Computationally, the procedure followed is to group the observations into classes of identical patterns of missing and observed components. Initial estimates of the mean and covariance matrix are obtained using (3.2.4) and (3.2.5) on the complete vectors, if there are any. Then (3.2.9) is used to get initial estimates of the missing values and the completed data used in (3.2.4) and (3.2.5) to get new estimates of the parameters. Finally the covariance matrix is corrected for bias by adding quantities like $(\Sigma^{m,m})^{-1}$ as indicated by (3.2.8). If there are no complete vectors, it is proposed to use some good initial guess of the missing data and then to start the cycle with (3.2.4) and (3.2.5) as before. It should be noted that the theory of partitioned matrices is quite useful in reducing the amount of computation since the inverse of the quite large matrix $\Sigma_{0,0}$ can be easily expressed in terms of the submatrices of Σ^{-1}, as can $\Sigma_{m,0}\Sigma_{0,0}^{-1}$. (See Woodbury, M.A., "Inverting modified matrices," *Stat. Res. Group. Princeton, N.J. Memo.*, Vol. 42, 1950.)

The convergence is, in certain cases, quite slow and hence methods of speeding it must be used. It must also be remembered that the correct procedure to follow for any analysis, such as multiple regression, is to use the corrected covariance matrix and the mean as data, since using only the values predicted by (3.2.9) will give rise to a biased covariance matrix. The necessary theory is quite easy to work out and computer programs have been written. The available information for estimating the mean is most easily obtained by accumulating the portion $\Sigma_{0,0}^{-1}$ contained in the observed components.

3.3. *Example 3.* Mixture problems may be regarded as missing information problems by noting that the indicator variable $Z_{n,k}$ (which is 1 if the nth observation is in the kth class and zero otherwise) is missing. If this was available then the constrained log likelihood would be

$$L = \sum_{n=1}^{N} \sum_{k=1}^{K} Z_{n,k}[\log p_k + f_k(x_n|\theta_k)] - N\left(\sum_{k=1}^{K} p_k - 1\right); \tag{3.3.1}$$

from this we may obtain the score for p_k as

$$\text{Sco}\,(p_k|x_n, Z_n) = \text{Sco}\,(p_k|Z_n) = \sum_{n=1}^{N} \frac{Z_{n,k}}{p_k} - N, \tag{3.3.2}$$

whilst the score for θ_k is

$$\text{Sco}\,(\theta_k|x_n, Z_n) = \sum_{n=1}^{N} Z_{n,k} \frac{1}{f_k} \frac{\partial}{\partial \theta_k} f_k(x_n|\theta_k). \tag{3.3.3}$$

If there is no missing data these equations may be separated into K classes, one set for each class.

If we do not observe the $Z_{n,k}$, then the image of $(x_n, Z_{n,1}, \cdots, Z_{n,k})$ is just x_n and hence we must find $E(Z_{n,k}|x_n)$ which is the posterior probability

$$P[k|x_n] = \frac{p_k f_k(x_n|\theta_k)}{f(x_n|\theta)}, \tag{3.3.4}$$

where

$$f(x_n|\theta) = \sum_{k=1}^{K} p_k f_k(x_n|\theta_k). \tag{3.3.5}$$

The image scores are

$$\text{Sco}\,(p_k|X) = \sum_{n=1}^{N} \frac{p[k|x_n]}{p_k} - N \tag{3.3.6}$$

and

$$\begin{aligned} \text{Sco}\,(\theta_k|X) &= \sum_{n=1}^{N} P[k|x_n] \frac{1}{f_k} \frac{\partial}{\partial \theta_k} f_k(x_n|\theta_k) \\ &= \sum_{n=1}^{N} P[k|x_n]\, \text{Sco}_k\,(\theta_k|x_n). \end{aligned} \tag{3.3.7}$$

The following estimating equations are obtained

$$\hat{N}_k = \sum_{n=1}^{N} P[k|x_n] \tag{3.3.8}$$

and

$$\hat{p}_k = \frac{\hat{N}_k}{N}. \tag{3.3.9}$$

The information computations require expressions for the expected values of the second derivative of the likelihood functions. Although these expressions do not have a form which can be easily evaluated for the mixture of normals, or any other standard distribution, they are being recorded here for the sake of completeness:

$$-\frac{\partial^2 L}{\partial p_i \partial p_j} = \frac{1}{p_i p_j} \sum_{n=1}^{N} P[i|x_n]P[j|x_n], \tag{3.3.10}$$

$$J^*(p_i p_j) = N \int \left[\frac{f_i(x|\theta_i) f_j(x|\theta_j)}{f(x|\theta)}\right] dx = N J^*_{i,j}(\theta), \tag{3.3.11}$$

where $J^*_{i,j}(\theta)$ is the above integral,

$$-\frac{\partial^2 L}{\partial p_i \partial \theta_j} = \frac{1}{p_i} P[i|x_n]P[j|x_n] \operatorname{Sco}(\theta_j|x_n), \tag{3.3.12}$$

$$J^*(p_i, \theta_j) = N p_i \frac{\partial J_{i,j}(\theta)}{\partial \theta_j} \tag{3.3.13}$$

$$-\frac{\partial^2 L}{\partial \theta_i \partial \theta_j} = \sum_{n=1}^{N} P[i|x_n]P[j|x_n] \operatorname{Sco}_i(\theta_j|x_n) \operatorname{Sco}_j(\theta_j|x_n), \tag{3.3.14}$$

$$J^*(\theta_i, \theta_j) = N p_i p_j \frac{\partial^2 J^*_{i,j}(\theta)}{\partial \theta_i \partial \theta_j}. \tag{3.3.15}$$

The overall nonsingular (unconstrained) information matrix is

$$J^* = \begin{bmatrix} J^*(p, p) & J^*(p, \theta) \\ J^*(\theta, p) & J^*(\theta, \theta) \end{bmatrix} \tag{3.3.16}$$

and its inverse is V^*. We note that

$$J^* \begin{bmatrix} p \\ 0 \end{bmatrix} = \begin{bmatrix} e \\ 0 \end{bmatrix}, \tag{3.3.17}$$

so that

$$V^* \begin{bmatrix} e \\ 0 \end{bmatrix} = \begin{bmatrix} p \\ 0 \end{bmatrix} \tag{3.3.18}$$

and

$$[e \quad 0]\, V^* \begin{bmatrix} e \\ 0 \end{bmatrix} = 1 \tag{3.3.19}$$

where e is a vector of ones. Thus the final covariance matrix of the estimator is

$$V^* - \begin{bmatrix} p \\ 0 \end{bmatrix} [p^T \quad 0] = \frac{1}{N} \begin{bmatrix} V^*(p, p) - pp^T & V^*(p, \theta) \\ V^*(\theta, p) & V^*(\theta, \theta) \end{bmatrix}. \tag{3.3.20}$$

The properties of J^* and V^* discussed in the papers of Aitchison and Silvey [5], [6] on constrained maximum likelihood estimation are also shared by the approximating sums of partial derivatives.

3.4. *Example 4.* Consider now the problem of estimating the parameters of a mixture of multivariate normal populations assumed to have equal covariance matrices but different means. Suppose that there are K populations and that we sample N observations. Thus, the data matrix X would have consisted of K submatrices, of N_k observations from population k for $k = 1, \cdots, K$. Instead, we have the image Y which consists of N observations from the mixture. If we had considered Y to consist of a single population then we would have obtained

$$\text{Sco}\,(\bar{\mu} \mid Y) = \Sigma^{-1} \sum_{n=1}^{N} (Y_n - \bar{\mu}) \tag{3.4.1}$$

leading to

$$\bar{\mu} = \frac{1}{N} \sum_{n=1}^{N} Y_n \tag{3.4.2}$$

and

$$\text{Sco}\,(\Sigma^{-1} \mid Y) = -\tfrac{1}{2} N\Sigma + \tfrac{1}{2} \sum_{n=1}^{N} (Y_n - \bar{Y})(Y_n - \bar{Y})^T \tag{3.4.3}$$

leading to

$$\Sigma = \frac{1}{N} \sum_{n=1}^{N} (Y_n - \bar{Y})(Y_n - \bar{Y})^T. \tag{3.4.4}$$

However, regarding Y as a mixture, we may write the likelihood function as

$$L(Y \mid \mu, \Sigma) = \sum_{n=1}^{N} \left[\sum_{k=1}^{k} p_k f_k(X_n) \right] \tag{3.4.5}$$

$$\begin{aligned} \log L = \sum_{n=1}^{N} \log \Bigg[\sum_{k=1}^{K} p_k (2\pi)^{-p/2} |\Sigma|^{-1/2} \\ \exp\{-\tfrac{1}{2}(X_n - \mu_k)^T \Sigma^{-1} (X_n - \mu_k)\} \Bigg] \end{aligned} \tag{3.4.6}$$

$$= -\frac{Np}{2} \log 2\pi + \frac{N}{2} \log |\Sigma^{-1}| + \sum_{n=1}^{N} \log \left[\sum_{k=1}^{k} p_k \exp\{-\tfrac{1}{2}(X_n - \mu_k)^T \Sigma^{-1} (X_n - \mu_n)\} \right].$$

Due to the imposed restriction, that $\Sigma_{k=1}^{K} p_k = 1$, we are required to maximize the function

$$L^* = \log L - \lambda\left(\sum_{k=1}^{k} p_k - 1\right), \tag{3.4.7}$$

where λ is a Lagrangian multiplier.

Defining $\text{Sco}\,(\theta|X) = (\partial L^*/\partial X) \log L^*$ we obtain

$$\text{Sco}\,(p_k|Y) = \sum_{n=1}^{N} \frac{f_k(Xn)}{f(X_n)} - \lambda, \tag{3.4.8}$$

where

$$f(X_n) = \sum_{k=1}^{K} p_k f_k(X_n). \tag{3.4.9}$$

Multiplying by p_k and summing over k gives $\lambda = N$, hence

$$\sum_{n=1}^{N} \frac{f_k(X_n)}{f(X_n)} = N \qquad \text{for} \quad k = 1, \cdots, K. \tag{3.4.10}$$

If we introduce the posterior probability

$$P[k|X_n] = \frac{p_k f_k(X_n)}{f(X_n)} \tag{3.4.11}$$

and define

$$\hat{N}_k = \sum_{n=1}^{N} P[k|X_n], \tag{3.4.12}$$

we obtain

$$p_k = \frac{\hat{N}_k}{N} \tag{3.4.13}$$

and

$$\hat{\mu}_k = \sum_{n=1}^{N} P[k|X_n] \frac{X_n}{\hat{N}_k}. \tag{3.4.14}$$

Finally

$$\hat{\sigma}_{i,j} = \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{N} P[k|X_n](X_{n,i} - \hat{\mu}_{k,i})(X_{n,j} - \hat{\mu}_{k,j}). \tag{3.4.15}$$

If we call this estimate of Σ the within class covariance matrix and denote it by $\Sigma^{(w)}$, then we may write

$$N\Sigma^{(w)} = \sum_{n=1}^{N} \sum_{k=1}^{K} P[k|X_n](X_n - \bar{\mu} + \bar{\mu} - \mu_k)(X_n - \bar{\mu} + \bar{\mu} - \mu_k)^T \tag{3.4.16}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} P[k|X_n][(X_n - \bar{\mu})(X_n - \bar{\mu})^T + (X_n - \bar{\mu})(\bar{\mu} - \mu_k)^T + (\bar{\mu} - \mu_k)(X_n - \bar{\mu})^T + (\bar{\mu} - \mu_k)(\bar{\mu} - \mu_k)^T]$$

$$= \sum_{n=1}^{N} (X_n - \bar{\mu})(X_n - \bar{\mu})^T - \sum_{n=1}^{N} \sum_{k=1}^{K} P[k|X_n](\bar{\mu} - \mu_k)(\hat{\mu} - \mu_k)^T.$$

Thus we may write

$$\Sigma^{(w)} = \Sigma^{(T)} - \Sigma^{(B)}, \tag{3.4.17}$$

where $\Sigma^{(T)}$ is the total variance and $\Sigma^{(B)}$ is the between population variance.

The computational procedure proposed is to take some good guess as an initial set of estimates for the parameters p_k, μ_k, Σ. Then to cycle iteratively through (3.4.14), (3.4.15), and (3.4.13) until the parameter estimates stabilize. At each stage we use the best parameter estimates available. As is usual with such iterative procedures the convergence may be slow and will require speeding in practice.

APPENDIX

Brief Review of the Literature and Historical Development

Although considered earlier by Allan and Wishart [7], the first general approach to the problem of missing data, in field experiments, was that of Yates [50], who provided formulae enabling the least squares estimate of a single missing datum to be computed from the row and column sums. He also provided a similar formula to correct for the bias introduced into the sums of squares, and suggested that the estimation formula could be used iteratively if there was more than one missing datum. However, no general correction for the bias in the sums of squares was given. The basic ideas behind the approach of Yates have been used by many authors since that time to cover most common linear models (see code U1). It has also been used for a factor analytic model (see Woodbury, Clelland, and Hickey [46], Woodbury and Siler [48]), but it is restricted to the univariate case with independent observations. A second approach to the problem was the "covariance method" (code U2) of Bartlett [11], (see also Coons [17]) which results in the same estimates as Yates but which also gives unbiased sums of squares. Tocher [39] described a method (code U3) which appears to be a combination of the approaches of Yates and Bartlett. This method gives biased sums of squares but a general correction is given. A final method (code U4) for univariate random variables is the iterative maximum likelihood method described by Hartley [28]. This involves replacing the missing values by their expected values, given the model and the parameters. Examples were given, by Hartley, for a number of discrete distributions having sufficient

statistics and for which the maximum likelihood parameter scores were linear in the observations.

The multivariate normal has been extensively studied; the first approach being the direct application of maximum likelihood (code M1). The first step was made by Wilks [45] who considered the general bivariate case. Special trivariate cases were considered by Lord [34] and Edgett [23]. Their approaches were greatly simplified by Anderson [9] who considered the general "nested" case for which the likelihood function could be factored. Trawinski and Bargmann [40] used the notation of Roy's general linear model to write the likelihood equations, and gave an example of a trivariate case for which each type of incomplete vector was observed an equal number of times. Another approach to the problem of testing hypotheses in such situations, where the components are missing by design rather than accident, is given by Kleinbaum [31]. Hocking and Smith [29] approached the problem by sequentially combining covariance estimators by adding one group at a time, starting with the complete observations. This method is statistically efficient and is the same as Anderson's maximum likelihood solution in the nested case. The above methods all appear to be valid approaches to the multivariate case. A method very similar to that of Hartley [28] was described by Federspiel, Monroe and Greenberg [24] (although used earlier by Greenberg). This is an iterative method which involves replacing the missing components by their conditional expectation, given the observed components. Although computationally simple, it gives rise to biased estimates of the covariance matrix (and, hence, of the mean), since the score contains quadratic functions of the observations. Buck [14] used a similar approach and also corrected for the bias in the covariance matrix, in the case of a single missing component. However, he failed to give the correct extension to more than a single missing component. Another approach (code M3), which could give rise to a nonpositive definite covariance matrix, is the use of all the available data to estimate each component of the mean and covariance matrix separately. This has been described by Glasser [25] and Haitovsky [27]. An extensive literative review was given by Afifi and Elashoff [2], who also considered some of the methods discussed here, as well as many of the methods in common use. One can determine from their extensive analysis that most approximations are best "quick and dirty" and at worst misleading.

Missing information problems, as distinct from missing data problems, are many and varied, some being recognized as such, some not. Examples of the lack of identifiability in mixture problems have been presented by Behboodian [12] for a mixture of univariate normal populations, by Wolfe [49] for a mixture of multivariate normal populations, and by Cohen [16] for the negative binomial, while the iterative methods used by Hartley [28] can deal, in addition, with the problem of lost information due to grouping, censoring and truncating. The problem, (code I2), of obtaining genotype frequencies from phenotype frequencies has been dealt with by Ceppellini, Siniscialco, and Smith [15].

REFERENCES

[1] A. A. AFIFI and R. M. ELASHOFF, "Missing observations in multivariate statistics I: Review of the literature," *J. Amer. Statist. Assoc.*, Vol. 61 (1966), pp. 595–604. (M2)

[2] ———, "Missing observations in multivariate statistics II: Point estimation in simple linear regression," *J. Amer. Statist. Assoc.*, Vol. 62 (1967), pp. 11–29. (M1)

[3] —— , "Missing observations in multivariate statistics III," *J. Amer. Statist. Assoc.*, Vol. 64 (1969), pp. 337–358.

[4] ———, "Missing observations in multivariate statistics IV," *J. Amer. Statist. Assoc.*, Vol. 64 (1969), pp. 358–365.

[5] J. AITCHISON and S. D. SILVEY, "Maximum likelihood estimates of parameters subject to restraints," *Ann. Math. Statist.*, Vol. 29 (1958), pp. 813–828.

[6] ———, "Maximum likelihood estimation procedures and associated tests of significance," *Ann. Math. Statist.*, Vol. 31 (1960), pp. 154–171.

[7] F. E. ALLAN and J. WISHART, "A method of estimating the yield of a missing plot in field experimental work," *J. Agric. Sci.: Camb.*, Vol. 20 (1930), pp. 399–406.

[8] R. L. ANDERSON, "Missing plot techniques," *Biometrics*, Vol. 2 (1946), pp. 41–47. (U1)

[9] T. W. ANDERSON, "Maximum likelihood estimates for a multivariate normal distribution when some observations are missing," *J. Amer. Statist. Assoc.*, Vol. 52 (1957), pp. 200–203. (M1)

[10] H. R. BAIRD and C. Y. KRAMER, "Analysis of variance of a balanced incomplete block design with missing observations," *Appl. Statist.*, Vol. 10 (1961), pp. 189–198. (U1)

[11] M. S. BARTLETT, "Some examples of statistical methods of research in agriculture," *J. Roy. Statist. Soc., Ser. B*, Vol. 4 (1937), pp. 137–183. (U2)

[12] J. BEHBOODIAN, "On a mixture of normal distributions," *Biometrika*, Vol. 57 (1970), pp. 215–217.

[13] J. D. BIGGERS, "The estimation of missing and mixed-up observations in several experimental designs," *Biometrika*, Vol. 46 (1959), pp. 91–105. (U1)

[14] S. F. BUCK, "A method of estimation of missing values in multivariate data, suitable for use with an electronic computer," *J. Roy. Statist. Soc., Ser. B*, Vol. 22 (1960), pp. 302–306. (M2)

[15] R. CEPPELLINI, M. SINISCIALCO, and C. A. B. SMITH, "The estimation of gene frequencies in a random mating population," *Ann. Hum. Genet.*, Vol. 20 (1955), pp. 97–115. (I2).

[16] A. C. COHEN, "A note on certain discrete mixed distributions," *Biometrics*, Vol. 22 (1966), pp. 566–571. (I1)

[17] I. COONS, "The analysis of covariance as a missing plot technique," *Biometrics*, Vol. 13 (1957), pp. 387–402. (U2)

[18] E. A. CORNISH, "The estimation of missing values in incomplete randomized block experiments," *Ann. Eugenics*, Vol. 10 (1940), pp. 112–118. (U1)

[19] ———, "The estimation of missing values in quasi-factorial designs," *Ann. Eugenics*, Vol. 10 (1940), pp. 137–143. (U1)

[20] C. W. COTTERMAN, "A weighting system for the evaluation of gene frequencies from family records," *Contributions from the Laboratory of Vertebrate Biology*, Ann Arbor, University of Michigan Press, 1947, pp. 1–21. (I2)

[21] D. B. DELURY, "The analysis of latin squares when some observations are missing," *J. Amer. Statist. Assoc.*, Vol. 41 (1946), pp. 370–389 (U1)

[22] N. R. DRAPER and D. M. STONEMAN, "Estimating missing values in unreplicated 2-level factorial and fractional factorial designs," *Biometrics*, Vol. 20 (1964), pp. 443–458. (U1)

[23] G. L. EDGETT, "Multiple regression with missing observations among the independent variables," *J. Amer. Statist. Assoc.*, Vol. 51 (1956), pp. 122–131. (M1)

[24] C. F. FEDERSPIEL, R. J. MONROE, and B. G. GREENBERG, "An investigation of some multiple regression methods for incomplete samples," *U. N. C. Inst. Statist. Memo*, No. 236, 1959. (M1), (M2), (M4)

[25] M. GLASSER, "Linear regression with missing observations among the independent variables," *J. Amer. Statist. Assoc.*, Vol. 59 (1964), pp. 834–844. (M3)

[26] W. A. GLENN and C. Y. KRAMER, "Analysis of variance of a randomized block design with missing observations," *Appl. Statist.*, Vol. 7 (1958), pp. 173–185. (U1)

[27] Y. HAITOVSKY, "Missing data in regression analysis," *J. Roy. Statist. Soc., Ser. B*, Vol. 30 (1968), pp. 67–82. (M3)

[28] H. O. HARTLEY, "Maximum likelihood estimation from incomplete data," *Biometrics*, Vol. 14 (1958), pp. 174–194. (U4)

[29] R. M. HOCKING and W. B. SMITH, "Estimation of parameters in the multivariate normal distribution with missing observations," *J. Amer. Statist. Assoc.*, Vol. 63 (1968), pp. 159–173. (M4)

[30] E. C. JACKSON, "Missing values in linear multiple discriminant analysis," *Biometrics*, Vol. 24 (1968), pp. 835–844. (M2)

[31] D. G. KLEINBAUM, "A general method for obtaining test criteria for multivariate linear models with more than one design matrix and/or incomplete response variates," *U.N.C. Inst. Statist. Memo*, No. 614, 1969. (M4)

[32] C. Y. KRAMER and S. GLASS, "Analysis of variance of a latin square design with missing observations," *Appl. Statist.*, Vol. 9 (1963), pp. 43–50. (U1)

[33] W. KRUSKAL, "The coordinate-free approach to Gauss-Markov estimation, and its application to missing and extra observations," *Proceedings of the Fourth Berkeley Symposium, Mathematical Statistics and Probability*, Berkeley and Los Angeles, University of California Press, 1961, Vol. 1, pp. 435–451. (U1)

[34] F. M. LORD, "Estimation of parameters from incomplete data," *J. Amer. Statist. Assoc.*, Vol. 50 (1955), pp. 870–876. (M1)

[35] G. E. NICOLSON, "Estimation of parameters from incomplete multivariate samples," *J. Amer. Statist. Assoc.*, Vol. 52 (1957), pp. 523–526. (M1)

[36] C. R. RAO, *Linear statistical inference and its applications*, New York, Wiley, 1965.

[37] ———, "Analysis of dispersion with incomplete observations on one of the characters," *J. Roy Statist. Soc., Ser. B*, Vol. 18 (1956), pp. 259–264. (M4)

[38] S. D. SILVEY, "The Lagrangian multiplier test," *Ann. Math. Statist.*, Vol. 30 (1959), pp. 389–407.

[39] K. D. TOCHER, "The design and analysis of block experiments," *J. Roy. Statist., Soc. Ser. B*, Vol. 14 (1952), pp. 45–100. (U3)

[40] I. M. TRAWINSKI and R. E. BARGMAN, "Maximum likelihood estimates with incomplete multivariate data," *Ann. Math. Statist.*, Vol. 35 (1964), pp. 647–657. (M4)

[41] A. WALD, "Tests of statistical hypotheses when the number of observations is large," *Trans. Amer. Math. Soc.*, Vol. 34 (1943), pp. 426–482.

[42] G. N. WILKINSON, "The analysis of covariance with missing data," *Biometrics*, Vol. 13 (1957), pp. 363–372. (U1)

[43] ———, "The estimation of missing values for the analysis of incomplete data," *Biometrics*, Vol. 14 (1958), pp. 257–286.. (U1)

[44] ———, "The analysis of variance and derivation of standard errors for incomplete data," *Biometrics*, Vol. 14 (1958), pp. 360–384. (U1)

[45] S. S. WILKS, "Moments and distribution of estimates of population parameters from fragmentary samples," *Ann. Math. Statist.*, Vol. 3 (1930), pp. 163–195. (M1)

[46] M. A. WOODBURY, R. C. CLELLAND, and B. J. HICKEY, "Applications of a factor analytic model in the prediction of biological data," *Behavioral Sci.*, Vol. 8 (1963), pp. 347–354. (M4)

[47] M. A. WOODBURY and V. HASSELBLAD, "Maximum likelihood estimates of the variance-covariance matrix from the multivariate normal," *SHARE National Meeting*, Denver, Colorado, March, 1970.

[48] M. A. WOODBURY and W. SILER, "Factor analysis with missing data," *Ann. New York, Acad. Sci.*, Vol. 128 (1966), pp. 746–754. (M4)

[49] J. H. WOLFE, "NORMIX—computational methods for estimating the parameters of multivariate normal mixtures of distributions," Technical Report, U.S. Naval Personnel Research Activity, San Diego, California, 1967, pp. 1–31.

[50] F. YATES, "The analysis of replicated experiments when field results are incomplete," *Emp. J. Expt. Agric.*, Vol. 1 (1933), pp. 129–142. (U1)

[51] F. YATES and R. W. HALE, "The analysis of latin squares where two or more rows, columns or treatments are missing," *J. Roy. Statist. Soc., Ser. B*, Vol. 6 (1939), pp. 67–79. (U1)